Modelos lineales generalizados

Javier Fernández-López, Profesor Ayudante Doctor
Unidad de Matemática Aplicada
Departamento de Biodiversidad, Ecología y Evolución, UCM

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

DOKTOREGO ESKOLA
ESCUELA DE DOCTORADO
DOCTORAL SCHOOL

Programas de doctorado:
- Agrobiología Ambiental
- Calidad y Seguridad Alimentaria
- Biodiversidad, Funcionamiento y Gestión de Ecosistemas

Bloque 2: Modelos lineales generalizados

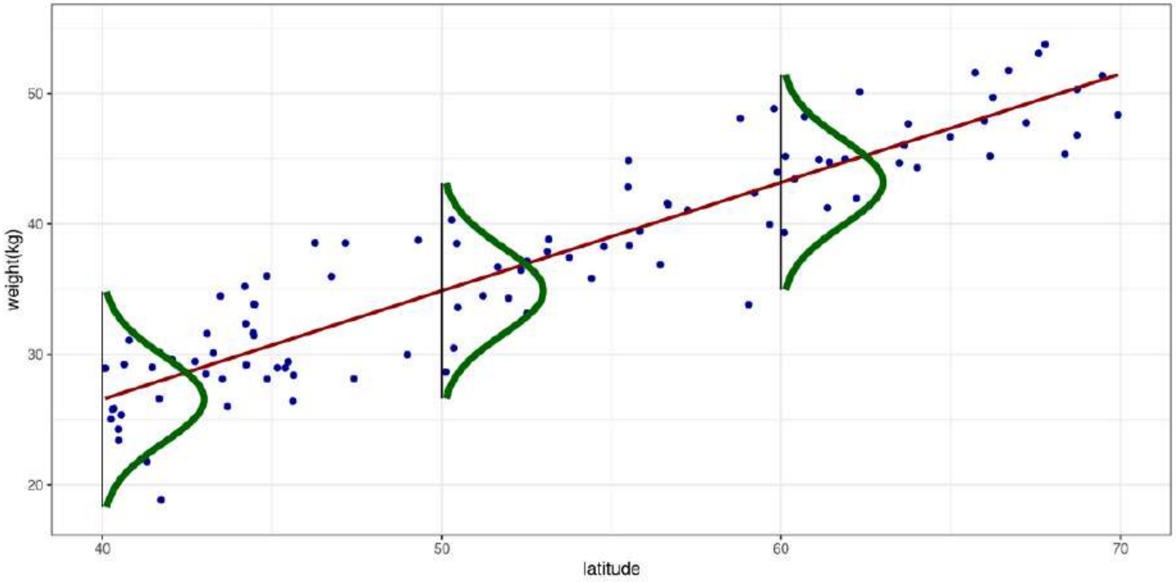
- Extendiendo el modelo lineal: factores categóricos
- Función vínculo: más allá de la distribución normal
- Evaluación y selección de modelos
- Predicciones

Factores de clasificación: predictores categóricos

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 \boxed{X_i}$$

Predictor cuantitativo continuo



Factores de clasificación: predictores categóricos

i	*concentrción*	*lote*
1	27	A
2	35	A
3	34	A
3	42	B
3	46	B
...	...	...
n	50	B

Factores de clasificación: predictores categóricos

i	*concentrción*	*lote*
1	27	A
2	35	A
3	34	A
3	42	B
3	46	B
...	...	...
n	50	B

$$y_i \sim \mathrm{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 \cdot \boxed{\mathrm{factor}_i}$$

Factores de clasificación: predictores categóricos

i	*concentrción*	*lote*	*dummy*
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...	...	...	...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 \cdot \boxed{\text{factor}_i}$$

$$\text{factor}_i = \begin{cases} 0, & \text{si el individuo } i \text{ pertenece al primer nivel del factor} \\ 1, & \text{si el individuo } i \text{ pertenece al segundo nivel del factor} \end{cases}$$

Factores de clasificación: predictores categóricos

i	*concentrción*	*lote*	*dummy*
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...	...	...	...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \boxed{\beta_0} + \cancel{\beta_1 \cdot \text{factor}_i} \longrightarrow 0$$

$$\text{factor}_i = \begin{cases} 0, & \text{si el individuo } i \text{ pertenece al primer nivel del factor} \\ 1, & \text{si el individuo } i \text{ pertenece al segundo nivel del factor} \end{cases}$$

Factores de clasificación: predictores categóricos

i	*concentrción*	*lote*	*dummy*
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...	...	...	...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \boxed{\beta_0 + \beta_1 \cdot \text{factor}_i}$$

1

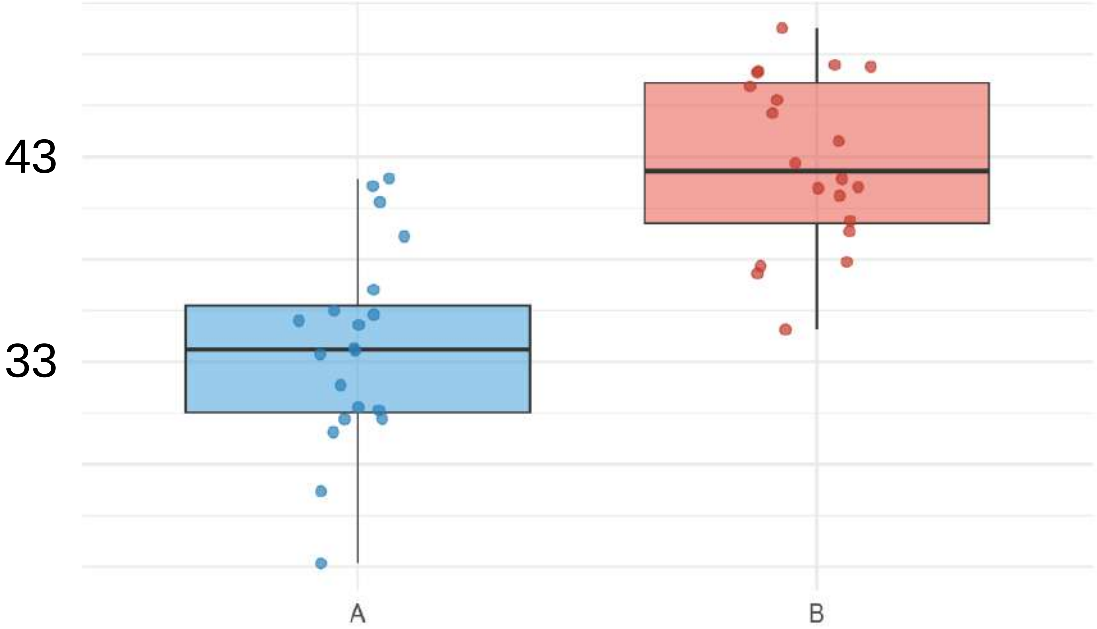
$$\text{factor}_i = \begin{cases} 0, & \text{si el individuo } i \text{ pertenece al primer nivel del factor} \\ 1, & \text{si el individuo } i \text{ pertenece al segundo nivel del factor} \end{cases}$$

Factores de clasificación: predictores categóricos

i	*concentrción*	*lote*	*dummy*
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...	...	...	...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 \cdot \text{factor}_i$$

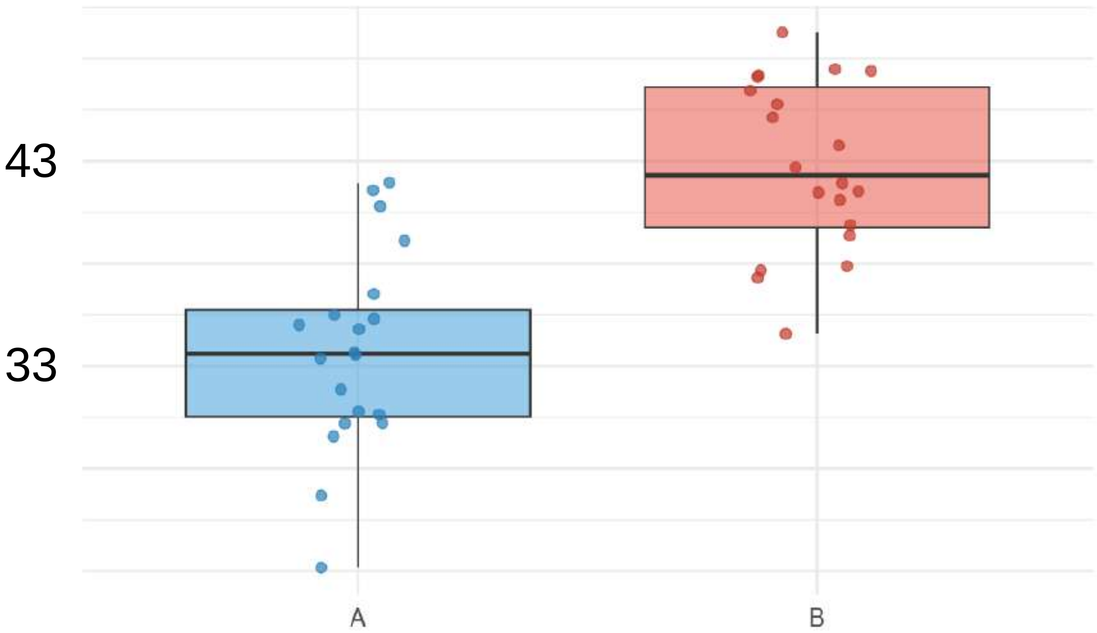


Factores de clasificación: predictores categóricos

i	*concentrción*	*lote*	*dummy*
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...	...	...	...
n	50	B	1

$$y_i \sim \mathrm{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 \cdot \mathrm{factor}_i$$

ANOVA



Caso práctico: fungicidas triazoles en heces de perdiz

Los fungicidas triazoles son compuestos químicos que se aplican habitualmente en semillas de cultivos para prevenir el crecimiento de hongos patógenos de plantas. Sin embargo, cuando las semillas son consumidas por la fauna silvestre, estos compuestos pueden producir efectos crónicos perjudiciales en su sauld y desarrollo. Queremos estudiar el efecto de los fungicidas triazoles sobre la condición corporal (peso) en perdices rojas (*Alectoris rufa*). Para ello se han capturado un total de 300 perdices en tres hábitats diferentes (semiurbano, agrícola y monte matorralizado) a las que se les ha sexado y extraido muestras de heces para obtener la concentración de fungicidas triazoles (ngramos de compuesto/gramo de heces).

Caso práctico: fungicidas triazoles en heces de perdiz

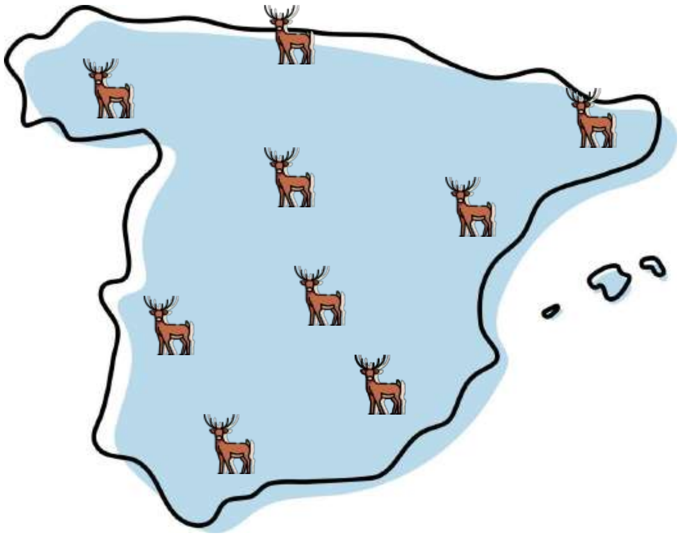
- Pensar variables: respuesta y predictores. Tipos de variables

- Construir un modelo desde lo sencillo a lo complejo

- Gráficos diagnóstico y bondad de ajuste

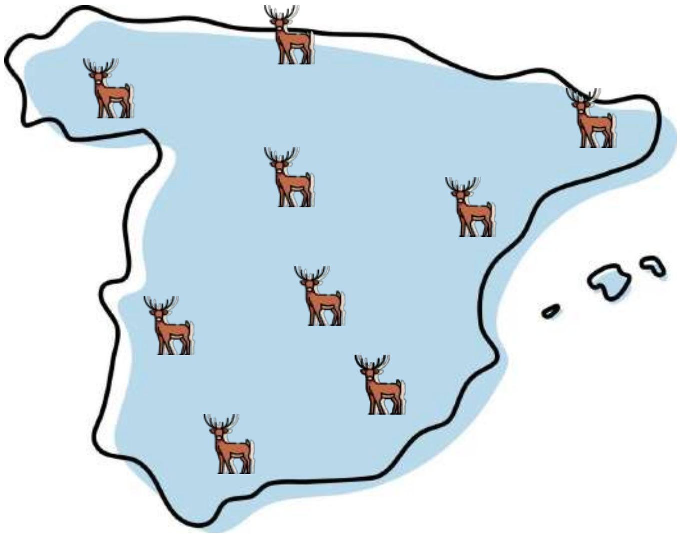
Distribución de la variable respuesta

Hasta ahora hemos utilizado la distribución Normal o Gaussiana para modelizar nuestra variable respuesta, pero no siempre es la más adecuada…



Distribución de la variable respuesta

Hasta ahora hemos utilizado la distribución Normal o Gaussiana para modelizar nuestra variable respuesta, pero no siempre es la más adecuada…



- N.º de agresiones entre gallinas dependiendo del fenotipo.

- Riqueza de sp. de musgos en función de diferentes variables edáficas.

- N.º de germinados/plántulas dependiendo de la influencia de una presa.

Distribución de la variable respuesta

Hemos muestreado 500 cuadrículas de 1x1km contando los ciervos que hemos visto en cada una de ellas… Creemos que el número de ciervos ya a estar relacionado con la temperatura en cada uno de los sitios de muestreo.

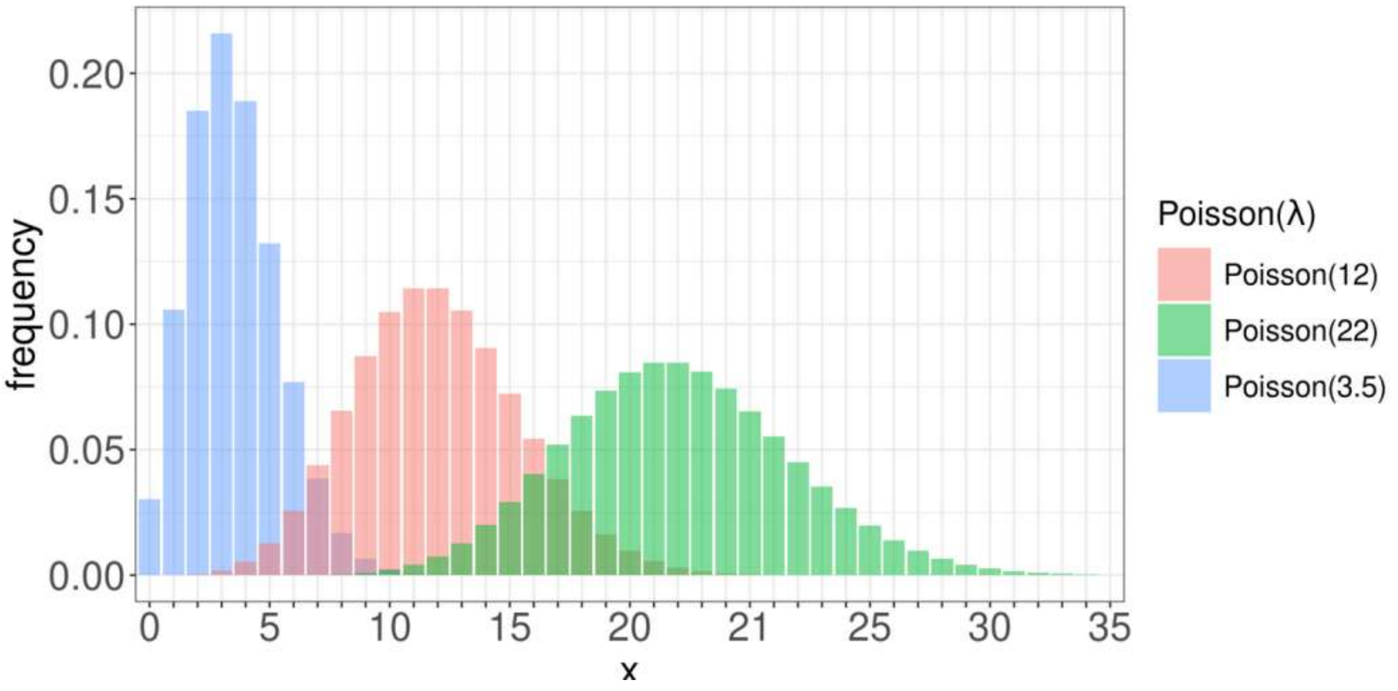
Si quisiéramos simular esta situación, ¿qué distribución de probabilidad podríamos utilizar?

Distribución de la variable respuesta

$$X \sim Poisson(\lambda)$$

Conteos

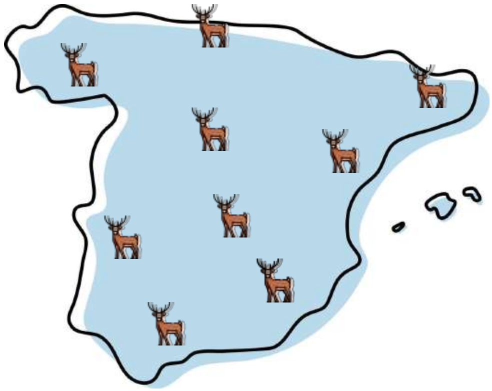
Support: $X \in (0, \infty)$ (natural); $\lambda \in (0, \infty)$ (real)



Distribución de la variable respuesta: Poisson

$$N_i \sim Poisson(\lambda_i)$$

$$\lambda_i = \beta_0 + \beta_1 Temperatura_i$$

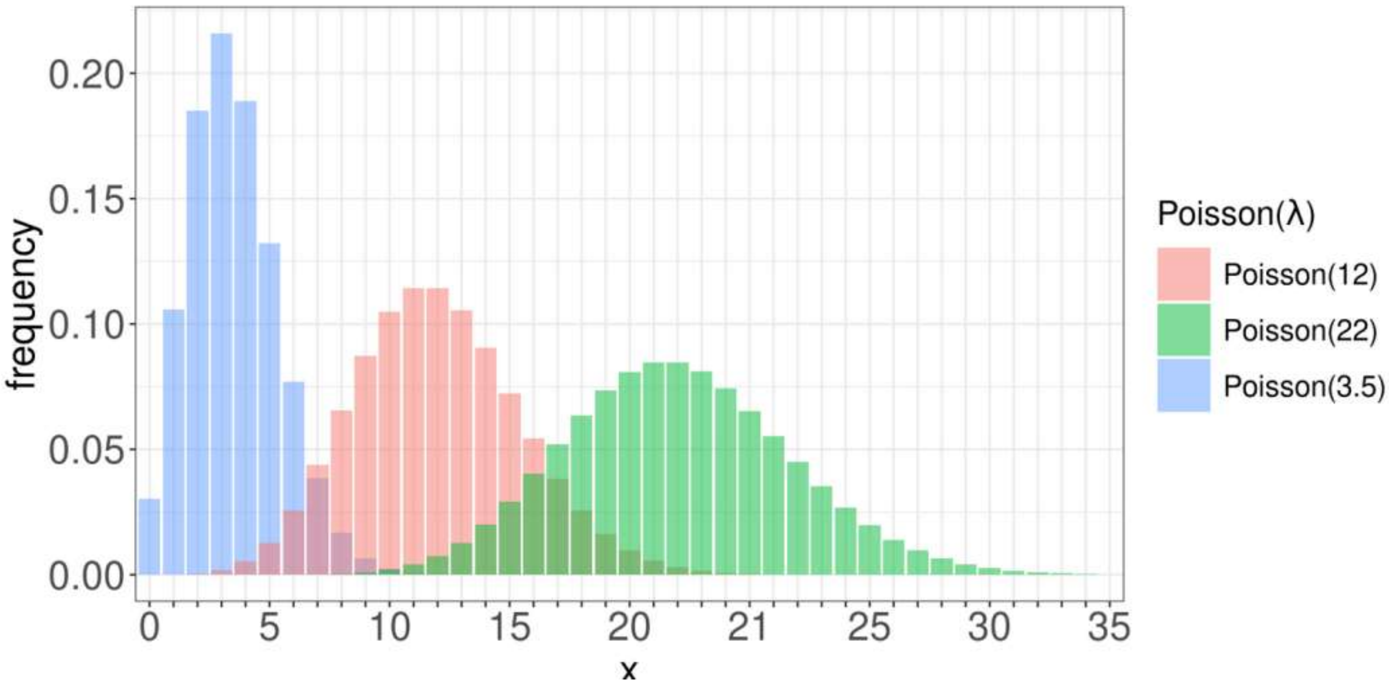


Distribución de la variable respuesta: Poisson

$$X \sim Poisson(\lambda)$$

Conteos

Support: $X \in (0, \infty)$ (natural) $\boxed{\lambda \in (0, \infty) \text{ (real)}}$



Distribución de la variable respuesta: Poisson

$$N_i \sim Poisson(\lambda_i)$$

$$\cancel{\lambda_i = \beta_0 + \beta_1 Temperatura_i}$$

$$log(\lambda_i) = \beta_0 + \beta_1 Temperatura_i$$

Funciones vínculo

Podemos usar cualquier otra distribución utilizando la misma maquinaria, simplemente se necesitarán funciones vínculo (links) diferentes para unir las covariables con parámetros de cada distribución: log, logit, etc.

Distribución	Links	Fórmula	Inversa
Gaussiana(μ, σ)	**Identidad**	$\mu = \beta X$	$\mu = \beta X$
Poisson(λ)	**Log**	$log(\lambda) = \beta X$	$\lambda = e^{(\beta X)}$
Bernoulli(*p*)	**Logit**, probit, cloglog	$log(\frac{p}{1-p}) = \beta X$	$p = \frac{e^{(\beta X)}}{1+e^{(\beta X)}}$
Binomial(*n*,*p*)	**Logit**, probit, cloglog	$log(\frac{p}{1-p}) = \beta X$	$p = \frac{e^{(\beta X)}}{1+e^{(\beta X)}}$

https://jabiologo.github.io/web/tutorials/prob_dist_links_merged.pdf

Con vuestros datos...

- Estudia tu variable respuesta. ¿Qué tipo es? Continua, categórica, admite decimales, valores negativos… ¿Puedo hacer un histograma con ella? ¿Qué pinta tiene?

- ¿Que distribución de probabilidad es la más conveniente para mi variable respuesta?

- Si es posible, elige dos predictores, uno continuo y otro categórico (en el caso de no disponer de esos tipos, elige dos predictores cualesquiera). Examina los predictores (histogramas, boxplot, gráficos de dispersión frente a la variable respuesta).

- Ajusta 4 modelos: el modelo nulo (~1), uno con cada variable y uno con ambas.

- Estudia: gráficos diagnóstico, p-valores, coeficientes para cada predictor, y devianza.